TTT/RR/## (do not delete this)

Using the TeraGrid for NOAA Scientific Computing

FY 2003 Proposal to the NOAA HPCC Program

August 19, 2002

| Title Page | Proposed Project | Budget Page |

Principal Investigator: **Tom Henderson**

Line Organization: OAR
Routing Code: R/FS5
Address:

Forecast Systems Laboratory
Advanced Computing Group
325 Broadway
Boulder, CO 80303-3328

Phone: (303) 497-6060
Fax: (303) 497-6301
E-mail Address: hender@fsl.noaa.gov

Dr. Albert J. Hermann
hermann@pmel.noaa.gov

Mark Govett
govett@fsl.noaa.gov

Dan Schaffer
schaffer@fsl.noaa.gov

Proposal Theme: **Next Generation Internet (NGI)**

Funding Summary: FY 2003 $ 80,600

Tom Henderson	Alexander E. MacDonald
Computer Scientist	Director
Forecast Systems Laboratory	Forecast Systems Laboratory

Using the TeraGrid for NOAA Scientific Computing

Proposal for FY 2003 HPCC Funding

Prepared by: Tom Henderson

Executive Summary:

We will investigate how NOAA can begin to make use of emerging Grid computing technologies. By combining geographically distributed computing resources into a single virtual resource, Grid computing will allow existing resources to be used more efficiently to meet NOAA's scientific challenges. It also will make it possible to combine resources to solve entirely new scientific problems, as has been demonstrated in other fields. We will begin by coupling simple component applications across multiple machines connected by the NSF-sponsored TeraGrid. We will explore performance issues such as allocation of resources to each component, coupling frequency, and Grid-related variations in run-times. We will also add "grid-awareness" to our existing application framework and make it available to all NOAA users. If this effort is successful, we will propose extending it to couple the Weather Research and Forecast Model (WRF) with the Regional Ocean Modeling System (ROMS) across multiple TeraGrid machines in FY 2004.

Problem Statement:

The accuracy and skill of atmospheric and oceanic prediction (forecasts, hindcasts and nowcasts) are limited by the resolution of appropriate physics in numerical models, the ability of such models to effectively assimilate data, and the availability of such data. Recent efforts to improve numerical models have included increasing spatial and temporal resolutions, extending spatial domains, running longer simulations, running ensembles, and coupling models that simulate coupled processes (such as ocean and atmosphere). Sensitivity studies, with different forcings and alternate parameterizations of unresolved physics, also entail a large number of simulations. These efforts are limited by the availability of adequate computing resources (for example, adequate to produce forecasts in a reasonable time). Unfortunately, computing resources are not presently used as efficiently as possible. One source of inefficiency is the difficulty of shifting workload from one machine to another due to geographic separation, different CPU architectures, different operating systems, and many other technical problems. Analogous to the modern electrical power distribution grid, Grid computing is a proposal to create a huge virtual supercomputing resource by connecting geographically distributed machines via a very high-speed WAN and providing a software "middleware" layer to insulate users from machine- and site-specific issues.

Grid-based computing has long been an attractive idea but until recently did not have sufficient network performance to run atmospheric and oceanic models efficiently across geographically distributed machines. The recently built TeraGrid has sufficient network performance. TeraGrid

is a multi-year effort to build and deploy the world's largest, fastest, most comprehensive, distributed infrastructure for open scientific research. When completed, it will include 13.6 teraflops of Linux cluster computing power distributed at four TeraGrid sites, facilities capable of managing and storing more than 450 terabytes of data, high-resolution visualization environments, and toolkits for grid computing. These components will be tightly integrated and connected through a network that will initially operate at 40 gigabits per second and later be upgraded to 50-80 gigabits/second: 8 times faster than today's fastest research network.

The TeraGrid project has connected large commodity Linux clusters at the National Center for Supercomputing Applications (NCSA), the San Diego Supercomputing Center, and other sites. Efforts in other scientific fields have already succeeded in making applications "grid-aware" so they can make use of the TeraGrid. NOAA must develop expertise in using Grid computing technology to take advantage of this new and expanding resource and to make better use of its own existing resources. Success in this area will contribute directly to the NOAA HPCC program's objective of enabling more accurate representation of the atmosphere-ocean system.

There are several ways the TeraGrid could be used to improve NOAA modeling. These include running a single model across multiple platforms to increase available memory and processors, running ensembles (multiple single-platform runs with slightly varying initial conditions or parameterizations), and running coupled models across multiple platforms. The NCSA's MEAD project (Modeling Environment for Atmospheric Discovery), is a multi-institutional effort to explore the usefulness of running ensembles of a coupled weather and ocean model (WRF and ROMS) on the TeraGrid. PMEL and FSL will both participate in MEAD. During the MEAD effort, each coupled model simulation will run on a single TeraGrid compute platform. The TeraGrid infrastructure will be used to launch hundreds of simulations simultaneously and will aid in storing and organizing large volumes (10's to 100's of terabytes) of resulting data.

Proposed Solution:

We propose to investigate the feasibility of running coupled models across multiple TeraGrid compute platforms. During FY 2003, we will focus on gaining access to the TeraGrid, developing expertise in using the system, and developing and coupling simple prototype components to demonstrate the feasibility of the approach. FSL will address all issues related to the TeraGrid and to distributed computing. PMEL will provide expertise in model coupling to ensure that the prototype components simulate the coupling characteristics of real models as closely as possible. We plan to leverage the MEAD effort by using the coupled WRF-ROMS as a guide. The prototype components will be used to measure the performance characteristics of the Grid including communications and I/O operations. In contrast to standard parallel computing environments in which the processors, memory, and networks are fairly homogeneous, the performance of Grid applications will be affected by the presence of nodes with different processor types and memory sizes and by an unbalanced communications network comprised of different bandwidth and latency characteristics. As part of this work, we will investigate the capabilities of the TeraGrid job launcher / queueing system, exploring the features that control the allocation of resources for each prototype component. If this effort succeeds in demonstrating the feasibility of running coupled models across multiple TeraGrid compute

platforms, PMEL's role will be expanded in a FY 2004 proposal to couple WRF and ROMS across multiple TeraGrid compute platforms.

High level software called the Scalable Modeling System (SMS), developed at FSL, will be leveraged in this effort. SMS is a directive-based parallelization framework for shared and distributed memory computers and has been used to parallelize many weather and ocean models, including ROMS. These models are used at domestic and foreign research institutions including several NOAA laboratories. Models parallelized using SMS run on most high performance computing platforms including IBM SP, SGI Origin, Cray T3E, Alpha-Linux clusters, and Intel-Linux clusters. The SMS parallel ROMS model is currently being run by PMEL on FSL's Alpha-Linux and Intel-Linux clusters. SMS will be extended to use communications packages such as the Globus-based implementation of MPI that already runs on the TeraGrid. FSL will provide matching base funds to support extension of SMS.

Some basic steps in this work include:

1. Investigate "middleware" packages for running applications on the TeraGrid such as Globus and MPICH-G2.
2. Run simple MPI-based codes on the TeraGrid using Globus and MPICH-G2 or other middleware.
3. Select the most mature middleware package(s) and integrate into SMS.
4. Develop very simple prototype components and couple them. Use the coupling of WRF and ROMS in the MEAD effort as a guideline to make the coupling communications between the prototype components as realistic as possible.
5. Use SMS to add parallelism and grid-awareness to the prototype components and test on one of FSL's Linux clusters.
6. Test prototype components coupled with one component running on FSL's Alpha-Linux cluster and the second running on FSL's Intel-Linux cluster.
7. Test prototype components coupled across the TeraGrid.
8. Investigate performance issues that arise including:
    a. Impacts of bandwidth and latency on coupling frequency.
    b. Performance variability due to time-variance of TeraGrid bandwidth and latency
    c. Methods of optimally assigning TeraGrid resources to each coupled component.
    d. Possible benefits of using data compression techniques to reduce communication bandwidth requirements.
    e. Performance tradeoffs between available bandwidth (which is proportional to the number of compute nodes used on the TeraGrid) and computational efficiency (which decreases as more compute nodes are used).
    f. Practical limitations on coupling frequency and the amount of data exchanged during coupling.
    g. Methods to mitigate communication latency.
9. Transfer experience gained to other NOAA laboratories via publication and release of grid-aware SMS.
10. Prepare for coupling of WRF and ROMS across multiple TeraGrid machines in FY 2004.

Analysis:

The main computing resources currently connected to the TeraGrid are Linux clusters built with commodity processors. FSL has acquired extensive experience using Linux clusters with different CPU architectures and different operating systems during the past three years. PMEL and FSL have a history of collaboration developing the parallel SMS ROMS model and running it on Linux clusters at FSL, on high-end workstations at PMEL, and on other platforms including IBM SP-2 and SGI Origin3000. Due to the MEAD effort, PMEL is uniquely positioned to become NOAA's first effective Grid user. This HPCC proposal effectively leverages these efforts to make Grid computing technology available for NOAA coupled modeling research with minimal costs.

In making the decision to pursue multi-platform coupling on the TeraGrid, we discarded two alternative approaches. We considered the alternative approach of testing a single atmospheric model distributed across multiple TeraGrid platforms. We chose to examine multi-platform coupling first because of the opportunity to leverage our effort in the MEAD project, and because we expect performance demands on the TeraGrid's high-speed WAN to be less for coupled models. The frequency of communication can be set relatively low for coupled ocean and atmosphere components, as compared to the frequency of individual time steps within either model. Further, the datasets communicated (e.g. sea surface temperature, surface air temperature, shortwave radiation, surface wind speed and direction) are small, relative to the memory requirements of either model. Hence, we expect performance will not be limited significantly by high network latency. We also considered running ensembles across the TeraGrid, but chose not to do so to avoid duplicating the MEAD effort.

Computing resources available on Grids such as the TeraGrid are huge and growing rapidly. If the effort proposed here is successful it will give NOAA scientists the ability to begin using TeraGrid resources to conduct their research. Ultimately, Grid computing technology could lead to more efficient use of existing NOAA computing resources. For example, NOAA might reduce overall IT costs by building its own computational Grid to permit efficient sharing of geographically distributed computational resources among the participating agencies.

Performance Measures:

This project will be successful if the prototype components can be coupled across multiple TeraGrid machines and if the prototypes are sufficiently similar to real oceanic and atmospheric models to allow simple extension of the techniques developed to benefit a coupled WRF-ROMS model.

Milestones

Month 1 -	Gain access to the TeraGrid.
Month 2-3 -	Investigate "middleware" packages for running applications on the TeraGrid such as Globus and MPICH-G2. Test simple MPI codes on the TeraGrid.

Month 4 -	Develop very simple prototype components and couple them. Test serial versions on a single process.
Month 4-5 -	Select the most mature middleware package(s) and integrate into SMS. Install middleware on FSL's Linux clusters to test new SMS features.
Month 6-7 -	Use SMS to add parallelism and grid-awareness to the coupled prototype components. Test on one of FSL's Linux clusters.
Month 8 -	Test prototype components coupled across two of FSL's Linux clusters using installed middleware.
Month 9 -	Test prototype components coupled across two TeraGrid machines.
Month 10-12 -	Investigate performance issues on the TeraGrid. Tune SMS as needed.
Month 12 -	Deliver final report. Release grid-aware SMS.

Deliverables

- Working simple coupled prototype components sufficient for examining how TeraGrid performance characteristics would affect performance of a real coupled ocean-atmosphere system such as WRF-ROMS.
- Report detailing the results of this performance study.
- Public-domain release of the Scalable Modeling System (SMS) software containing support for Grid computing.

Administrative Officer:	Sandra Aschert
Phone:	(303) 497-6803
E-mail Address:	Sandra.J.Aschert@noaa.gov

FMC Number: 947